# BIT at TREC 2010 Blog Track: Faceted Blog Distillation

Peng Jiang[1], Qing Yang[1], Chunxia Zhang[2], Zhendong Niu[1]

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China
[2] School of Software, Beijing Institute of Technology, Beijing 100081, China
{jp, cxzhang, yangqing2005, zniu}@bit.edu.cn

**Abstract.** This paper presents the work done for the TREC 2010 faceted blog distillation task. As the approach used in TREC 2009, a mixture of language models based on global representation is employed to rank the entire blogs by relevance and facets. The parameters in our approach are adjusted according to the experimental results in TREC 2009. In addition, we make use of the results evaluated in TREC 2009 to train a SVM classifier. This classifier is used to filter and re-rank the results obtained by the mixture model.

## 1 Introduction

This is the second year that Beijing Institute of Technology (BIT) participates in TREC faceted blog distillation. We evaluated the effectiveness of combining different language models for faceted blog ranking and employing global representation for feed distillation [1]. The parameters in our approach are tuned to the best performance according to the experimental results in TREC 2009. In addition, we make use of last year's evaluated results to train a SVM classifier model, followed by filtering and re-ranking results obtained by the mixture model. With the improved approach and more integrated indexed data (In TREC 2009, the index for our submitted results does not contain the data of January 2008) experimental results have shown the significant improvement this year.

## 2 Our approach for Faceted Blog Distillation Task

The goal of faceted blog distillation task is to find blogs that are principally devoted to a given topic and a given facet. The retrieval unit is a blog but a single document. All retrieved blogs should not only be relevant to the given topic, but also to the given facet, such as opinionated, personal and in-depth facets. According to the approaches used in last year, this task will be run as two separate sub-tasks, namely the Baseline Blog Distillation and the Faceted Blog Distillation.

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **NOV 2010** | | **00-00-2010 to 00-00-2010** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **BIT at TREC 2010 Blog Track: Faceted Blog Distillation** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Beijing Institute of Technology,School of Computer Science and Technology,Beijing 100081, China,** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES
**Presented at the Nineteenth Text REtrieval Conference (TREC 2010) held in Gaithersburg, Maryland on 16-19 November 2010. The conference was co-sponsored by the National Institute of Standards and Technology (NIST), the Defense Advanced Research Projects Agency (DARPA), and the Advanced Research and Development Activity (ARDA).**

14. ABSTRACT
**This paper presents the work done for the TREC 2010 faceted blog distillation task. As the approach used in TREC 2009, a mixture of language models based on global representation is employed to rank the entire blogs by relevance and facets. The parameters in our approach are adjusted according to the experimental results in TREC 2009. In addition, we make use of the results evaluated in TREC 2009 to train a SVM classifier. This classifier is used to filter and re-rank the results obtained by the mixture model.**

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT **unclassified** | b. ABSTRACT **unclassified** | c. THIS PAGE **unclassified** | **Same as Report (SAR)** | **6** | |

## 2.1 System Overview

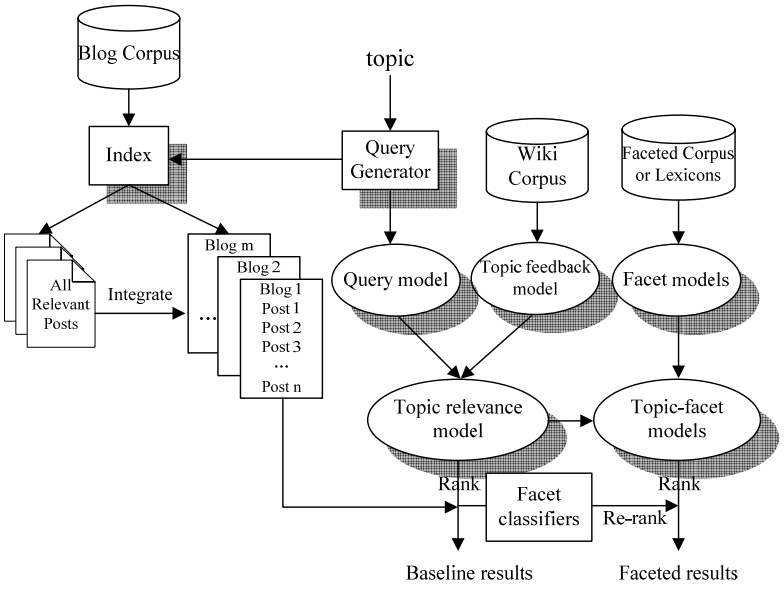


**Fig. 1. The System Architecture.**

As shown in Fig.1, our system consists of four main parts: query generator, blog retrieval component, language model building component and classification component. Query generator is responsible for parsing topic and query generation. The blog retrieval component indexes blog corpus, and finds all blog posts corresponding to each query. All posts are then combined to blogs according to their feedno. After this, the component retrieves all posts that have the same feedno. The language model building component is responsible for generating topic-facet mixture models. The topic-facet mixture model is a linear combination of topic relevance model and facet model. In order to build topic relevant language model, we index Wikipedia corpus[1], and finds relevant wiki pages about a given topic. The classification component is in charge of filtering and re-ranking the results generated by the retrieval system using mixture models.

## 2.2 Query Generation

This year, we only choose title field of a topic for query string generation. According to the experiments on TREC 2009, we notice that using title, description and narrative fields (TDN) cannot achieve a better performance than using title field (T) alone.

[1] Wikipedia corpus is obtained from ClueWeb09 Dataset (Category B subset).

### 2.3 Dataset and Index

TREC Blog08 collection contains permalinks, feed xml and homepages. We use the permalinks and feed xml for the faceted blog distillation task. The feed xml corpus is used to train the facet classifiers. The feed xml file format is similar to that of permalinks file, but the content is encoded by xml. We modified some Indri code to index the feed xml corpus. Krovetz stemmer and a list with 450 stop words (e.g. a, about, above or many other common but useless words) are used.

### 2.4 Retrieval Model

We choose Global Representation Model to represent blog as last year. This model treats a blog as a virtual document which is comprised of all posts of the blog. Thus, this model can effectively reflect the recurring interest in a given topic over the time span of the feed. In addition, since we use language model based approach to rank feed, Global Representation model, which combines many posts into a large document, can avoid the problem of sparsity of words as far as possible. In our experiments in TREC 2009, we notice that global representation model can achieve a better performance than pseudo-cluster selection model mentioned in previous work[2].

In order to combine all posts into one large virtual document, first find all relevant posts for a given topic, which are then combined into some feeds by their feedno fields. After this, metadata search is used to collect all posts of a given feed by the feed's feedno. Finally, we obtain all relevant feeds which contain not only relevant posts but irrelevant posts as well.

We rank feeds by the Kullback-Leibler Divergence of a feed language model and a topic-facet language model. In this solution, two different language models are defined: one for a topic and its' facet value ($\theta_{TF}$); and another for the virtual document of a feed ($\theta_D$). That is, we assume that $\theta_{TF}$ represents the topic and facet information need, while $\theta_D$ represents a feed. The KL-divergence of these two models is able to measure how close they are to each other. Thus, the distance can be used to rank feeds:

$$score(D,TF) = \sum_w p(w \mid \theta_{TF}) \log p(w \mid \theta_D) + cons(\theta_{TF}) \quad (1)$$

Because the constant $cons(\theta_{TF})$ (the entropy of $\theta_{TF}$) does not affect the results of ranking feeds, we do not compute it in our system. Thus, the main task is to estimate $\theta_{TF}$ and $\theta_D$. $\theta_D$ can be estimate by the following formula:

$$p(w \mid \theta_D) = \frac{c(w,D) + \mu p(w \mid C)}{|D| + \mu} \quad (2)$$

where $p(w|C)$ is a background language model, $C(w,D)$ is the count of $w$ occurs in $D$, and $\mu$ is a Dirichlet smoothing parameter. We use $\mu$=2000, which is optimal in most case[3]. The $\theta_{TF}$ in equation (1) is the language model which reflects not only the topic information need but facet information need. So we use a mixture of language model to estimate it. We assume the feed is generated from a mixture of topic relevant model $\theta_T$ and facet model $\theta_F$. The topic-facet model $\theta_{TF}$ is a linear combination of $\theta_T$, and $\theta_F$:

$$\theta_{TF} = (1-\beta)\theta_T + \beta\theta_F \quad (3)$$

where $\beta$ is used to control the influence of the facet model $\theta_F$. In equation (3), $\theta_T$ is the topic relevance model that can be obtained by pseudo-relevance feedback method (PRF).

Using Formula (1), we can obtain a ranked list for each facet. Then we use last year's results to train three facet classifiers. These classifiers are used to filtered some results with low facet confidence (according to the classification confidence) in the ranked list. Finally, we merge the rank score and classification confidence by the ratio according to experiments in TREC 2009.

## 3 Experimental Results

We submit 2 baseline runs and 14 faceted runs. Table 1 shows the results of our two baseline runs and three standard baseline results. Table 2 shows the overview of our all faceted runs results, best and median results over all submitted runs. Table 3 shows the improvement over our results in TREC 2009.

**Table 1. Baseline Results**

| Run id | Map | R-prec | bpref | P@5 | P@10 |
|---|---|---|---|---|---|
| BITblog10bl1 | **0.3519** | **0.3845** | **0.3332** | **0.5500** | **0.4750** |
| BITblog10bl2 | 0.3025 | 0.3562 | 0.3059 | 0.4667 | 0.4000 |
| stdbaseline1 | 0.3210 | 0.3749 | 0.3249 | 0.5083 | 0.4250 |
| stdbaseline2 | 0.2117 | 0.2613 | 0.2057 | 0.3333 | 0.2958 |
| stdbaseline3 | 0.1832 | 0.2625 | 0.2021 | 0.3667 | 0.3292 |

**Table 2. Faceted Results (Map)**

| Run id | all | opinion | factual | indepth | shallow | official | personal |
|---|---|---|---|---|---|---|---|
| BIT10bl1fd1 | 0.1946 | 0.1590 | 0.2423 | 0.2232 | 0.1436 | 0.2239 | 0.1834 |
| BIT10bl1fd2 | 0.2023 | 0.1748 | 0.2717 | 0.2232 | 0.1436 | 0.2384 | 0.1849 |
| BIT10bl1fd3 | **0.2367** | 0.2129 | **0.3920** | 0.**2232** | 0.1436 | 0.3411 | 0.1779 |
| BIT10bl1fd4 | 0.2222 | 0.1636 | 0.3885 | 0.1515 | 0.1607 | 0.3510 | 0.1803 |
| BIT10bl2fd1 | 0.1714 | 0.1868 | 0.2369 | 0.1814 | 0.124 | 0.1745 | 0.1651 |
| BIT10bl2fd2 | 0.1801 | 0.2278 | 0.2513 | 0.1814 | 0.124 | 0.1888 | 0.1663 |
| BIT10bl2fd3 | 0.2084 | **0.2395** | 0.3468 | 0.1976 | 0.1173 | 0.2833 | 0.1529 |
| BIT10bl2fd4 | 0.1845 | 0.2021 | 0.2687 | 0.1814 | 0.124 | 0.2492 | 0.1341 |
| BIT10std1fd1 | 0.2013 | 0.1968 | 0.2553 | 0.2232 | 0.1436 | 0.2092 | 0.2056 |
| BIT10std1fd2 | 0.2094 | 0.224 | 0.2825 | 0.2232 | 0.1436 | 0.2242 | 0.2037 |
| BIT10std1fd3 | 0.2002 | 0.2032 | 0.2663 | 0.2232 | 0.1436 | 0.2152 | 0.1843 |
| BIT10std1fd4 | 0.1925 | 0.2009 | 0.2508 | 0.2232 | 0.1436 | 0.1597 | **0.2086** |
| BIT10std2fd1 | 0.1246 | 0.1054 | 0.2341 | 0.1146 | 0.1084 | 0.1228 | 0.1056 |
| BIT10std3fd1 | 0.1091 | 0.0886 | 0.1217 | 0.0951 | 0.0856 | 0.1689 | 0.0951 |
| BITblog10bl1 | 0.2219 | 0.1636 | 0.3870 | 0.1529 | **0.1607** | **0.3487** | 0.1803 |
| BITblog10bl2 | 0.1836 | 0.1982 | 0.2707 | 0.1804 | 0.1236 | 0.247 | 0.1341 |
| median | 0.1043 | 0.1373 | 0.1806 | 0.0725 | 0.1012 | 0.0894 | 0.1693 |
| best | **0.4411** | **0.4941** | **0.5512** | **0.2557** | **0.4247** | **0.4299** | **0.5606** |

Figures 2 and 3 show our system's per-topic performance for the faceted blog distillation in terms of MAP, alongside with the per-topic median and best performance over all groups' runs, respectively. All topics are sorted along the *x* axis in descending order of BIT best run performance.

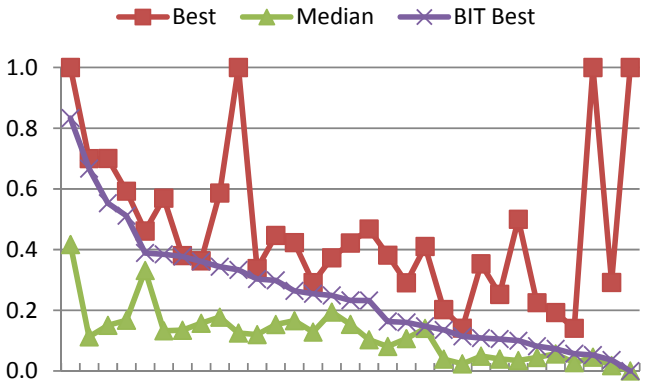


**Fig. 2. MAP of faceted blog distillation for the first faceted value**

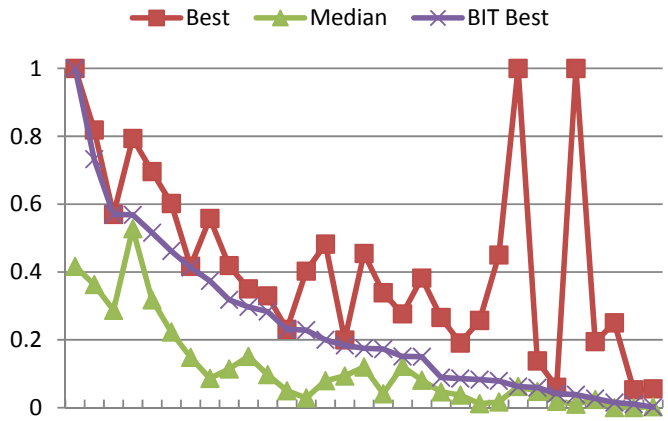


**Fig. 3. MAP of faceted blog distillation for the second faceted value**

**Table 3. Improvement over last year's results**

| Run | Baseline result | Faceted result |
|---|---|---|
| BIT 09 best run | 0.1165 | 0.1026 |
| BIT 10 best run | **0.3519** | **0.2367** |
| improvement | 202.06% | 130.07% |

## 4 Conclusion

From the experimental results, we notice a significant improvement over last year's result, although overall performances of all other groups' approach are improved. In fact, the basic retrieval model, that is topic-facet mixture model, is similar to the model used in TREC 2009. In TREC 2009, if the missing index of blogs in January 2008 is added, we can achieve relatively low but similar experimental results. A significant difference is parameters adjustments according to the results in TREC 2009. Another difference is the use of SVM classifier for filter and re-ranking.

**Acknowledgments.** This work is supported by the grant from Chinese National Natural Science Foundation (No: 60705022) and Beijing Institute of Technology Excellent Doctoral Dissertation Yu Miao Foundation.

## References

1. Peng Jiang, Qing Yang, Chunxia Zhang, Zhendong Niu, *BIT at TREC 2009 Faceted Blog Distillation Task*, in The Eighteenth Text REtrieval Conference (TREC 2009) Proceedings. 2009: NIST.
2. J. Seo and W. B. Croft, *UMass at TREC 2007 Blog Distillation Task*, in The Eighteenth Text REtrieval Conference (TREC 2007) Proceedings. 2007: NIST.
3. Chengxiang, Z. and L. John, *A study of smoothing methods for language models applied to information retrieval.* ACM Trans. Inf. Syst., 2004. 22(2): p. 179-214.